Historical Software Issue 12: Statistical Package for the Social Sciences/SPSS X

Thaller, Manfred

software

ISSUE 10: STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES / SPSS X

GENERAL CONSIDERATIONS

SPSS is certainly the most widely applied system of statistical analysis within historical-social research. It also is a rather old system, which did not change many features of design since its introduction in the late sixties. As a consequence during the last years many criticisms have been raised against its use, alleging, that the system did in many respects not represent any more the state of the art of statistical software.

SPSS Inc. took up this challenge some years ago by a double strategy: a number of statistical procedures were added to the basically unchanged package, while at the same time, parallely to the distributed version of the system, another one has been developed which was built around a completely new concept of files and a new operational logic. This version is now being distributed as SPSS X, being accompanied at the same time by a completely new documentation(1).

"SPSS X is both a data management and analysis program." (Guide 257) This sentence probably describes best the intentions behind the re-design. The package was heavily criticized (as almost all the statistical ones) for its strict adherence to the rectangular data matrix and even more so for certain clumsy features in data manipulation. As a result the new version of SPSS consists of the almost unchanged statistical modules of version 9, embedded into a data management system. (This change having been completed, SPSS Inc. now is of course working to increase the number of the statistical methods available, see eg. KW 34 2.)

Besides this basic enhancement of the scope of the system, numerous minor refinements have been applied. The most obvious is that the time honoured column sixteen has disappeared: every line which has a blank in column one is now considered a continuation line. Laudable (though this reviewer never understood, why it took so long to incorporate this obvious improvement) as this change is, one might consider as an even greater pleasure the general tendency to reduce the amount of default output produced.

The streamlined statistical system, embedded into a data management environment, is clearly an answer to the threat posed by SAS to the market position of SPSS. This is not only obvious by some of the capabilities described below, but even more so by part of the documentation and by some of the plans announced: SAS provides documentation on how to use the system for the analysis of commonly occuring types of data sources - the new SPSS documentation does the same (see Census and "SMF", a volume dealing with

Address all communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge-Weg 11, D 34 Göttingen

how SPSS may be used to analyse the System Maintenance Facilities data of IBM computers). SAS has a MATRIX manipulation language - SPSS announced one as an immediate aim for the further development. SAS provides very good information and a number of tools for the integration of a users own routines into the system - SPSS which had, · years ago, such facilities, described in the first edition of the manual, promises to reintroduce them soon (Guide ii).

Besides this defensive measures against a strong competitor, particularly on the commercial and semicommercial (e.g. medical) markets, SPSS still has a different policy than SAS: we are never confronted with any claim, that the system might be a Higher Programming Language. Indeed, the scope of the canned solutions for data management goes (as far as such comparisons can reliably be made) further than the ones provided by SAS - if you reach their limit however, you are right back to writing your own FORTRAN, PL/I or whatever programm. Similiarly, despite the announcement, that the user would soon be able to incorporate his/her own routines written in a higher programming language, SPSS X is still shielded very well against any attempt to use it within one project together with other packages, using the highlights of one to overcome the shortcomings of the other: SCSS (the conversational product of SPSS Inc.) is the only system, towards which an interface is provided (2).

Every producer of software has nowadays to face the micro challenge. SPSS Inc. takes it up by offering two products: SPSS/PC, being based on SPSS X for use on the IBM PC, and SPSS/Pro, being based on the "old" SPSS and intended for the Digital Professional 350 (See KW 33 2-3). If this policy - to provide a different version of the system for every brand of micro - is very wise, the future will have to show. This reviewer has to admit, that, having had his nerves injured already by SPSS-11, a highly praised version of SPSS for PDP's 11 series, he would like to work with those micro products before commenting upon them.

a. STATISTICAL METHODS AVAILABLE

As an example for the general layout of the command language we can use the command to compute a correlation coefficient.

```
PEARSON CORR AGE INCOME
OPTIONS 5
```

will compute Pearson correlation coefficients between the variables AGE and INCOME, in the currently active file (on this concept see below letter "d"), suppressing the printing of the number of cases and significance levels.

So, besides the possibility to forget about column sixteen, nothing has changed? Not quite. While the "old" commands (roughly: the ones available with version 6) still employ the old logic "use the command to specify what you want to have computed, use OPTIONS to control missing values and the layout of print, use STATISTICS to define which coefficients you want to have computed/printed" the commands introduced since then use OPTIONS and STATISTICS either not at all or only as subsidiary means, while most of the specification is done with keywords embedded into the specification as "subcommands", as LOG in:

```
BOX-JENKINS VARIABLE=WHEATPRICE/LOG/IDENTIFY
```

which would calculate the natural logarithm of WHEATPRICE before plotting

the autocorrelation function. (An exception is the FREQUENCIES command: the former OPTIONS have been integrated as subcommands into the FREQUENCIES specification, the STATISTICS card remains, but expects no numbers (as all the other ones), but keywords as a specification.)
We already mentioned, that the statistical procedures are basically the same as in level 9 of "old" SPSS.

The commands for frequency counts and simple distributions of variables (Guide 264 - 285) have been overhauled with the intention of providing additional coefficients (including arbitrary quantiles, which Im glad about for teaching purposes and Lorentz curves), at the same time reducing the number of coefficients being printed without being asked for. We finally have not to explain to beginners any more what the kurtosis is! Last, but not least, the output is nowadays much prettier and more close to being reproducible at least for conference papers. CONDESCRIPTIVE allows now with an elegant construct of the language to compute and store in the systemfile the z-scores with one command - a feature we will comment upon later.

CROSSTABS and BREAKDOWN (Guide 286 - 331) have been overhauled as well with the same tendencies, but with less obvious changes: SPSS annnounces the intention to include a TABLE BUILDER within one of the next releases (Guide ii) for more refined tables, so we will probably be left with CROSSTABS operating under the already mentioned logic "get less without asking, be able to ask for more" for quite some time for research purposes, while having at some time in the future the TABLE BUILDER at our disposal which is to complex to handle for day to day work, while producing output beautiful enough for publication.

If this TABLE BUILDER can be handled by the general user of the computer out of historical-social research, I have quite a few doubts: SPSS has featured since quite some time a REPORT procedure, which allows to design quite sophisticated reports which could be very well applied to historical micro analysis (nominal lists with the average income, taxes and so on of the people within a historical community) but has to the best of my knowledge scarcely be used so, being controlled by a syntax considerably more complex and much harder to use than the remainder of the commands (Guide 332-375). Historical-social researchers seem not to have been the only users frightened away by this syntax, as SPSS Inc. sees the need to announce a new and - according to the promise - vastly improved REPORT BUILDER (Guide ii) on top of the current REPORT facility.

The T-TEST procedure is basically unchanged (Guide 431-436).

The analysis of variance (Guide 439-462) is still being supported by the old and relatively powerless procedures ANOVA and ONEWAY, while with the more recent MANOVA procedure (multivariate analysis of variance and covariance program, Guide 464-539) a rather powerful mechanism for general linear models exists. (As of release 2 of SPSS X, available that far only on IBM operating under OS or CMS and on VAX computers, LISREL is said to be available as an option additionally. No information on the features supposed to be released by this version, beyond their being mentioned in KW 34, p 2. was available for this review.)

Regression analysis is supported by a new REGRESSION procedure that is vastly superior to the old one in release 6 and has been introduced already earlier (Guide 601-622). Correlation analysis is annoyingly still split between a procedure for PEARSON CORRelation (Guide 579-588) and another one for

NONPARamatric CORRelation (Guide 663-670). Partial Correlation (Guide 589-600), Discriminant Analysis (623-646) and Factor Analysis (Guide 647-662) remain more or less unchanged. Practically without importance for historical-

social research, is testing for the reliability of complex scales (Guide 717-734).

This reviewer thinks however, that one should point at the so far scarcely noticed potential of some of the other acquisitions of SPSS during the last few releases which have gone unnoticed by many users, as no integrated manual was available. This is - on the low end of complexity - certainly true for the MULT RESPONSE procedure (Guide 303-320) which could solve quite a few of the problems arising out of people having two jobs at the same time, being found in a source with more than one place of "origin" and so on. This is in principle true for the nonparametric tests (Guide 671-696) which would make sense with the vast load of historical data, where no presuppositions regarding the statistical distribution are possible, and even the analysis of survival rates (guide 735-748) holds ome untapped potential for historical demography.

Besides these procedures for statistical analysis, which could be used by followers of historical-social analysis, there are three which are specifically useful for the field: since a number of releases Time Series Analysis is supported (Guide 697-716). This procedure includes only the more simple varieties of Time Series Analysis, specifically not Spectral Analysis. I admit, however, that I still have too many doubts about the quality of historical data particularly favoring Spectral Analysis, that I could see a very large shortcoming in that.

The only statistical addition to SPSS X against release 9 of the "old" system is very important for the readers of this newsletter: SPSS X incorporates Loglinear Analysis (Guide 541-570). I feel not able to evaluate the statistical merit of this implementation: as my experience with the method goes (dating a couple of years back and acquired with the help of ECTA), I think, however, that the procedure should be sufficiently comfortable to be handled without to much risk of falling into traps by historians.

Release 2 of SPSS X is supposed to contain additionally a simplified procedure for Loglinear Analysis, Logistic Regression and "hierarchical clustering for a small to moderate number of cases". All of these techniques are very interesting for historical-social researchers, but, nothing known about the procedures to this reviewer besides their supposed existence (KW 34 p. 2), it is hard to comment upon them.

b. LEARNING AND TEACHING

The documentation of SPSS X is good(3). It contains already very useful introductory volumes which are separate from the main manual, being split between a cartoon-illustrated introduction into the EDP side of SPSS (Basics) and an introductory volume into the statistical background of the procedures (Statistics). Welcoming that from a didactical point - the prospective user is not awed by 800 pages at the very beginning - this reviewer thinks it somewhat dangerous for the traditionally not statistically overeducated historical-social researchers to have the statistical explanations taken away from the potentially misapplied technical descriptions.

Its data manipulation language definitely beyond the flexibility of SAS, SPSS

X can even less than this package claim to be one for teaching. statistics. The control language is very useful for getting even complex results quickly, you cannot use it, however, to show to the beginner which computations are behind the results displayed.

c. ENHANCED DATA HANDLING CAPABILITIES

The first thing one notices looking upon the new descriptions of the data transformation commands is the same general streamlining we already noticed in our comments upon the statistical procedures: RECODE now has a keyword INTO, which allows one to forget about the COMPUTE needed previously to keep the unchanged value, so

```
COMPUTE      CHANGEDA=A
RECODE       A (1,2,3,4=101) ...
```

becomes

```
RECODE A (1,2,3,4=101) ... INTO CHANGEDA
```

The cumbersome LAG command has been turned into a function, so the terrible construct

```
LAG          LAGA5,LAGA4,LAGA3,LAGA2,LAGA1=LAG
```

needed to get a five-step lagging variable becomes simply

```
COMPUTE LAGA5=LAG(A,5)
```

A (relatively moderate) number of additional functions for statistical purposes, random number generation and similar things are provided. Useful: "logical" functions to abbreviate cumbersome expressions.

```
IF           (A EQ 1 OR 102 OR 203 OR 304 OR 51)
```

becomes

```
IF ANY(A,1,102,203,304,51)
```

A much more important change: SPSS redesigned the concept of MISSING VALUES completely. As before one has the possibility to define a value as missing; once this is done, however, the system takes care, that an additional "system missing value" will be assigned to all variables derived out of computations which encountered variables with a missing value. This is an extremely valuable feature and the differentiation between a user-defined missing value and a system defined one (accompanied by the introduction of a whole set of related functions for using missing values in logical expressions) seems to this reviewer to be much more convincing, than the idea to recode (and loose by that) internally all missing value codes to a system standard value. As in other cases one should mention however, that by this concept SPSS has made good ground lost during recent years, when compared to other packages - certainly not broken new one.

Very important is another conceptual change. SPSS differentiates now between the input file, the (stored) system file and the currently active file. Every job is at every point of time always related to one active file. During any job this relation can be changed an arbitrary number of times. So at any point of time data can be added from a second, third, ... nth systemfile to

the one already worked upon. Besides another streamlining effect (no DELETE VARS/KEEP VARS "immediately before SAVE", but a KEEP=varlist/DROP=varlist' clause with any command that makes a systemfile active or turns the active file into a systemfile) this means, that it is not necessary any more, to differentiate overly sharply between temporary and permanent data modifications. Permanent data modificiations are possible at every point of a job, and any part of the data modifications can be declared to be temporary by a new command. (Additionally the concept of a "scratch variable" was introduced, for variables needed just intermediately during computations: no discovery any more, that you forgot to include the variables DUMMY1 TO DUMMY100 on your DELETE VARS command, blowing up your systemfile by that.)

The most obvious expansion of the data modification commands: DO IF ... END IF ... ELSE IF constructs and LOOP controls are provided. While not as sophisticated as SAS in this respect, you get enough tools to perform incomparably more complex data modifications than so far. As with SAS (which was obviously copied in some respects) you can use these tools together with a set of specialized ones within an INPUT PROGRAM ... END INPUT PROGRAM construction which allows complex input formats beyond the ones provided by the standard commands. A very nice idea: there is an inbuilt check for the number of times a loop is executed (which can be influenced by the user) - no red heads of beginners any more, looking at their printout having been covered with 10.000 repetitions of the same line by an endless loop.

The old SUBFILE concept has been dropped. YOU can SPLIT FILE your data now, which has a similiar effect, but can be administrated much more elegantly. The same shortcoming has to be noted as with SAS: in this and other cases you have to SORT your data explicitly, the system being unable to trigger this by itself.

The labeling features remain more or less unchanged, with the exception of a very agreeable redisign of the syntax: all labels are included between apostrophes or quotation marks, putting an end to a whole number of ambiguities in the use of certain keywords and characters.

The feature I like most: we already mentioned, that you can very easily produce z-scores with the help of the CONDESCRIPTIVE command. Im still thinking about, how I can get it into the head of beginners, that z-scores have a mean of 0 and a standard deviation of 1 - but when this is overcome, they are a very elegant means of doing parameter dependent definitions of additional variables.

A USERPROC, allowing to include ones own programm, is announced and allegedly (KW 34, p.3) included already in the release 2 for the VAX. Due to the software design of the VAX it allows the use of almost all higher programming languages - what becomes of that, when the system is implemented on other mainframes, remains to be seen.

d. MORE COMPLEX FILE HANDLING

This is a highlight of SPSS X. While, due to its less flexible data modification language, it does not allow for the really complex file structures as SAS does, the system provides standard solutions for two of the most common problems of non-rectangular files, which can be used with much less experience than any SAS solution can.

The first is the case of files, which by various reasons shall be adminis-

trated as one file, but contain cases which have only a small number of variables (if any) in common. A typical example would be a file, where every case contains some information on a person, some of the cases followed by 5 or 10 variables describing a piece of land owned by this person, some of them followed by the description of a debtor, some by that of a creditor, others again by the description of some mobile property and so on. While in theory such a situation can simply be handled by inserting the data into different columns on the different cases, in practical work such an approach leads to numerous errors and to very large data files. Here SPSS X allows the definition of MIXED FILES which enable you to define an arbitrary number of different RECORD TYPES, each of which has its own input format, being identified by one variable which has to have the same position within all record types.

A second case, called somewhat preposterously a "file type" by SPSS would be the GROUPED FILE. Seen without the intention to sell it, it is a somewhat more humble thing: GROUPED FILES consist of cases of equal length and equal structure, which are spread across more than one line of input. Provided are quite good means to detect missing or misplaced lines during input.

Most important however is the third file type, the NESTED FILE: it provides for hierarchical structures, as the old problem of the family consisting of one household card and 20 cards describing the members of that household. One should have a few reservations about the efficiency of the administration of such structures by the standard means of SPSS X and Ifm afraid many enthusiastic users will find themselves thrown against the limitations of their computing centres disk storage - still, for many research projects this is a solution that is only slightly more complex to handle than the classical INPUT FORMAT card and incomparably more powerful.

Merging of data sets is supported very well. As in the case of SAS the only problem that remains is, what happens to variables which appear with different values in two files - but this is a limitation built into the very concept of the scalar variable.

A very nice side-effect of the new concept of the active file is a changed behaviour of the AGGREGATE command: when you execute AGGREGATE now, the aggregated file becomes the active file immediately, so cou can work upon it without defining a new systemfile. The difference between the "true" and "compositional" aggregated files has disappeared, as the merging routines now allow to add the aggregated information from one case of the aggregated file to as many cases of the original file as necessary. A happy by-note: the enourmous masses of almost perfectly useless printout produced by the old AGGREGATE seem not to be produced by SPSS X any more!

e. STRING HANDLING IN STATISTICAL PACKAGES

Inferior to SAS in this respect, SPSS X includes still the most important string handling subroutines as available in higher programming languages. Reading the manual, one has the impression, however, that when the string-handling features were provided, people concentrated decidedly upon the problems, posed by the administration of personal or place names, and even more so on the problems of short alphabetic abbreviations (for which special support exists). So, what happens to the system performance when variables going to the full maximum length of 255 (fixed length) characters are being handled in a complex way, remains to be seen.

f. IMPROVED DISPLAY FACILITIES

SPSS X contains SPSS Graphics consisting of "three procedures - PIECHART, BARCHART, and LINECHART - that are fully integrated into the SPSS X system". What this means is not perfectly clear, but, as it is well known, that the overpriced graphics option of SPSS 9 (virtually identical with the three routines in SPSS X) was rather a flop with unwilling buyers, this reviewer understands, that, as the price has risen considerably anyway, the graphics are included now.

The graphics produced are restricted to what the routines are named for - piecharts, barcharts and linecharts - and are pretty nice, controlling a sufficient number of fonts, shading colors and the like.

One problem can not be decided from the information available for this reviewer: It is well known, that SPSS Inc. developed its graphics options being heavily dependend on the software of ISSCO. (DISSPLA and TELL-A-GRAF.) Indeed an interface to TELL-A-GRAF is provided. The question is, how far the typical European computing centre will be able to provide for the necessary drivers for its plotting devices and how far the readiness to support this software will go. DISSPLA, being certainly a very good package, has known to be less successful on the European academic market recently, as it does not cling to the GKS standard which is favored - at least within Europe - nowadays.

SPSS X is bad on producing pretty diagrams on a line printer. (A promise of reform has been made). SCATTERGRAM remains unchanged and the (line printer) plotting facilities of the various statistical routines are as well.

EDA is basically unsupported. SPSS X is able, however, to produce "Box-and-Whisker-Plots" and "Stem-and-Leaf-Plots" on a line printer. Indeed it has been able to do so since a number of releases. Why SPSS Inc. decided to bury this feature in the depths of the description of the MANOVA procedure, without mentioning it even in the index, remains mysterious to this reviewer.

CONCLUSIONS

SPSS X has considerably more power than the product most of us have been using since some years. This reviewer supposes, that in about 50 % of the projects of historical-social research he knows, the effort to go into the definition of file structures and writing SPSS programs can be reduced dramatically by switching to SPSS X. Still, the system remains a closed one, without interfaces into other program systems and with data manipulation facilities inferior to the ones supported by SAS. On the other hand, SPSS X as a whole remains an easy-to-use package. There are few if any facilities one cannot expect a typical researcher in historical social studies to master.

So SPSS can have a considerable immediate impact upon current research projects. However, as the effort needed to move to the efficient use of its features from the previous standard could be quite large, as some of the central concepts have changed, we can recommend it unreservedly only in those cases, where the researcher is doing his programming all for himself. In research setups, where some expertise in computing is provided centrally for a project, one should at least look at the other packages.

SPSS X is a vast improvement over SPSS that far. It has a very comfortable

handling of the most frequent problems in handling of non rectangular files. In all other respects, it did not surpass its competitors, however: so when you have access to another package as well, think first, if, as you have to learn a number of new concepts anyhow, it might not be worth while to make a more drastic change as soon, as your computing centre stops the support of the old version of SPSS(4).

NOTES

1) The following volumes were used for this review: (short titles as used in in-text-quotations are underlined)
SPSS X Users Guide, New York etc.: McGraw-Hill, 1983.
SPSS X Basics, New York etc.: McGraw-Hill, 1984.
Marija J. Norusis: SPSS X Introductory Statistics Guide, New York etc.: McGraw-Hill, 1983.
The two last issues of the user bulletin of SPSS "KeyWords" (34, Winter 1984 and 33, Fall 1983) are quoted as KW 34 and KW 33 respectively.
The following volumes, being announced and shortly commented in the Guide, were not yet available at the time of this review:
Marija J. Norusis: SPSS X Advanced Statistics Guide.
SPSS X Analysis of United States Census Data.
SPSS X Analysis of SMF Data.
The following volume, being announced in the Guide and already described in a McGraw-Hill prospectus was, according to our dealer, withdrawn from the McGraw-Hill program and will therefore, if at all, only be published later:
SPSS X Data Management.
Additionally a new version of the volume "SPSS Statistical Algorithms" has been published. As it is of virtually no importance to any but the most refined users, we do not consider it here.

2) SPSS X contains a very nice IMPORT/EXPORT feature to move data between varying SPSS X implementations on different brands of computers.

3) This reviewer doesn't like reviews counting misprints. The Users Guide contains one of the most beautiful examples however he has seen so far: on page xi we are astonished to learn from the table of contents, that for the formulation of numeric expressions we have to take care of "The Order of nations". The diplomatic puzzle is resolved on page 91 - even analysing international summits, the computer still is only interested in the order of operations.

4) SPSS X is available with version 1 on most IBM systems, with version 2 on VAX computers. It is said to become available within the next weeks for Sperry (formerly UNIVAC) and some Honeywell Bull computers.